Multiple kernel multivariate performance learning using cutting plane algorithm

Jingbin Wang
National Time Service Center, Chinese Academy of Sciences, Xi' an 710600 , China
Graduate University of Chinese Academy of Sciences, Beijing 100039, China
jingbinwang1@outlook.com

Haoxiang Wang
Department of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14850, USA

Yihua Zhou
Department of Mechanical Engineering and Mechanics, Lehigh University, Bethlehem, PA 18015, USA

Nancy McDonald
Department of Computer Science, Tulane University, New Orleans, LA 70118, USA
nancya.mcdonald@yahoo.com

***Abstract*—In this paper, we propose a multi-kernel classifier learning algorithm to optimize a given nonlinear and nonsmoonth multivariate classifier performance measure. Moreover, to solve the problem of kernel function selection and kernel parameter tuning, we proposed to construct an optimal kernel by weighted linear combination of some candidate kernels. The learning of the classifier parameter and the kernel weight are unified in a single objective function considering to minimize the upper boundary of the given multivariate performance measure. The objective function is optimized with regard to classifier parameter and kernel weight alternately in an iterative algorithm by using cutting plane algorithm. The developed algorithm is evaluated on two different pattern classification methods with regard to various multivariate performance measure optimization problems. The experiment results show the proposed algorithm outperforms the competing methods.**

***Index Terms*—Pattern recognition, multiple kernel, multivariate performance measures, cutting plane algorithm**

I. Introduction

In different pattern classification problems, various performances are employed to evaluate the classifiers, including classification accuracy (ACC), F1 score , Matthews Correlation Coefficient (MCC), area under the receiver operating characteristic (ROC) Curve (AUC) and recall-precision break even point (RP-BEP) of recall-precision curve. Due to the nonlinear and nonsmooth nature of many performance measures, it is difficult to optimize them directly to learn an optimal classifier. To solve this problem, Joachims [1] proposed a support vector machine learning method for multivariate Performance measures (SVM^{Perf}). This other method has been applied to optimized some nonlinear multivariate performance measures to learn linear classifiers successfully. However, it is limited to the learning of linear classifiers. When data samples of different classes cannot be separated by a linear boundary, it is suggested to employ the kernel trick to map the data samples to a nonlinear high-dimensional data space so that a linear boundary could be learned [2], [3], [4]. Joachims and Yu [5] also extended the SVM^{Perf} to its kernel version to handle the nonlinearly distributed data. One important shortage of this method lies on the choosing of an optimal kernel function with its corresponding parameter. In [5], the RBF-Kernel is used to classification problems on some data sets without any justification, but it is highly doubt if this kernel is suitable for other data sets. Moreover, how the optimal parameter of the kernel function possibly influences the results significantly. One possible way to solve this problem is to conduct an exhausting linear search or a cross validation in the kernel function and parameter space by using the training set, which is very time-consuming and also makes the learned classifier over-fitting to the training samples.

To solve this problem, we assume that the desired kernel can be obtained by the linear combination of some candidate kernel functions with different kernel parameters. The optimal kernel is parameterized by the linear combination weights associated with different kernels. This framework is called Multi-Kernel Learning (MKL) since we explore the nonlinear kernel spaces of multiple kernels [6]. To learn the kernel weights, we cast the MKL problem with the multivariate performance measures problem, and proposed an unified learning problem for both MKL and multivariate performance measures problems. For the first time, we propose the problem of learning an optimal kernel for multivariate performance measures, and a novel solution for this problem by learning kernel in multiple kernel spaces simultaneously with optimizing multivariate performance measures.

The rest parts of this paper are organized as follows: in section II, we introduce the novel method by formulating the problem first, optimizing it then, and developing an iterative algorithm finally, in section III, the proposed method is evaluated on some benchmark data sets, and in section IV, the paper is concluded.

II. Proposed method

A. Problem Formulation

We assume we have a training data set with n training samples, and the training samples are organized in an training

matrix $X = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, where the i-th column $\mathbf{x}_i$ is the d-dimensional feature vector of the i-th training sample. Moreover, we also organize the class labels in a class label vector $\mathbf{y} = [y_1, \cdots, y_n]^\top \in \{+1, -1\}^n$, where $y_i \in \{+1, -1\}$ is the binary class label of the i-th training sample. Under the framework of kernel learning [7], an sample vector can be mapped into a high dimensional nonlinear Hilbert Space, via a implicit mapping function $\phi : \mathbf{x} \to \phi(\mathbf{x}) \in \mathcal{R}^{d'}$, where $d' \gg d$ is the dimension of the Hilbert Space. The mapping function is explored by a kernel function, which is defined as the dot-produce of the mapping of two samples $\mathbf{x}_i$ and $\mathbf{x}_j$, as $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$. In the multi-kernel learning framework, we may have several such Hilbert Spaces available and there corresponding nonlinear mapping functions are denoted as $\{\phi_m(\mathbf{x}) \in \mathcal{R}^{d'_m}\}_{m=1}^M$, where M is the number of Hilbert Spaces, $\phi_m(\mathbf{x})$ is the nonlinear mapping function of the m-th mapping function, and d'_m is the dimension of the m-th Hilbert Space. We also define the kernel function for the m-th Hilbert space as $K_m(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)_m^\top \phi_m(\mathbf{x}_j)$. We weight and concatenate the mapping function to form a longer vector in a more general Hilbert Space, $\phi_{\boldsymbol{\tau}}(\mathbf{x}) = \left[\tau_1 \phi_1(\mathbf{x})^\top, \cdots, \tau_M \phi_M(\mathbf{x})^\top\right]^\top \in \mathbb{R}^{d'}$ where $\tau_m \in \mathbb{R}_+$ is the nonnegative weight for the m-th Hilbert Space, $\boldsymbol{\tau} = [\tau_1, \cdots, \tau_M]^\top \in \mathbb{R}_+^M$ is the weight vector, and $d' = \sum_{m=1}^M d'_m$ is the dimension of the general Hilbert Space. Its corresponding kernel function is given as

$$K_{\boldsymbol{\tau}}(\mathbf{x}_i, \mathbf{x}_j) = \phi_{\boldsymbol{\tau}}(\mathbf{x}_i)^\top \phi_{\boldsymbol{\tau}}(\mathbf{x}_j) = \sum_{m=1}^M \tau_m^2 K_m(\mathbf{x}_i, \mathbf{v}_j) \quad (1)$$

It can be seen that the kernel function is also a weighted linear combination of the M kernel functions of the M Hilbert spaces. We map all the samples to the Hilbert spaces, and organize the mapping results in a $d' \times n$ matrix as $\phi_{\boldsymbol{\tau}}(X) = [\phi_{\boldsymbol{\tau}}(\mathbf{x}_1), \cdots, \phi_{\boldsymbol{\tau}}(\mathbf{x}_n)] \in \mathbb{R}^{d' \times n}$. We can also apply the kernel function to the matrix and obtain the $n \times n$ kernel matrix $K_{\boldsymbol{\tau}}(X, X) = \sum_{m=1}^M \tau_m^2 K_m(X, X) \in \mathbb{R}^{n \times n}$, where $K_m(X, X) = [K_m(\mathbf{x}_i, \mathbf{x}_j)] \in \mathbb{R}^{n \times n}$ is the kernel matrix of the m-th Hilbert space.

We consider the problem of learning a hypotheses function $h_{\mathbf{w}}(X)$ which maps a tuple of n samples organized in a data matrix X to a label vector of n labels $\mathbf{y}$. To this end, we first map the data matrix X to the general Hilbert space $\phi_{\boldsymbol{\tau}}(X)$, and then apply a linear discriminant function of the following form

$$h_{\mathbf{w}}(X) = \underset{\mathbf{y}' \in \{+1,-1\}^n}{\arg\max}\ \mathbf{w}^\top \phi_{\boldsymbol{\tau}}(X)\mathbf{y}' = \underset{\mathbf{y}' \in \{+1,-1\}^n}{\arg\max} \sum_{i=1}^n \mathbf{w}^\top \phi_{\boldsymbol{\tau}}(\mathbf{x}_i) y'_i \quad (2)$$

where $\mathbf{w} \in \mathbb{R}^{d'}$ is the parameter vector. Actually, it is equal to the following prediction results,

$$h_{\mathbf{w}}(X) = sign\left(\mathbf{w}^\top \phi_{\boldsymbol{\tau}}(X)\right) \quad (3)$$

where $sign(\cdot)$ is an element-wise $sign$ operation function.

To avoid the over-fitting problem, we try to reduce the complexity of the hypotheses function parameter $\mathbf{w}$ by minimizing the squared ℓ_2 norm,

$$\min_{\mathbf{w},\xi,\boldsymbol{\tau}} \left\{\frac{1}{2}\|\mathbf{w}\|_2^2 = \frac{1}{2}\mathbf{w}^\top \mathbf{w}\right\} \quad (4)$$

We also want to reduce the prediction error of the hypotheses function on the training set. To measure the prediction error, a loss function can be applied to compare the true class label tuple $\mathbf{y}$ against the output of the hypotheses function $h_{\mathbf{w}}(X)$. The following optimization problem is obtained with a $\Delta(\mathbf{y}, h_{\mathbf{w}}(X))$,

$$\min_{\mathbf{w}}\ \Delta(\mathbf{y}, h_{\mathbf{w}}(X)). \quad (5)$$

Instead of trying to optimize $\Delta(\mathbf{y}, h_{\mathbf{w}}(X))$ directly, we try to find its upper boundary and then minimize its upper boundary. Given (2), we have the following inequalities,

$$\begin{aligned} &\mathbf{w}^\top \phi_{\boldsymbol{\tau}}(X) h_{\mathbf{w}}(X) \geq \mathbf{w}^\top \phi_{\boldsymbol{\tau}}(X)\mathbf{y}', \forall \mathbf{y}' \in \{+1, -1\}^n \\ &\Rightarrow \Delta(\mathbf{y}, h_{\mathbf{w}}(X)) + \mathbf{w}^\top \phi_{\boldsymbol{\tau}}(X)\left(h_{\mathbf{w}}(X) - \mathbf{y}\right) \geq \Delta(\mathbf{y}, h_{\mathbf{w}}(X)) \end{aligned} \quad (6)$$

Thus we have the upper boundary of $\Delta(\mathbf{y}, h_{\mathbf{w}}(X))$, and the optimization problem in (5) can be relaxed to

$$\min_{\mathbf{w}}\ \left\{\Delta(\mathbf{y}, h_{\mathbf{w}}(X)) + \mathbf{w}^\top \phi_{\boldsymbol{\tau}}(X)\left(h_{\mathbf{w}}(X) - \mathbf{y}\right)\right\}. \quad (7)$$

We further relax the minimization of $\Delta(\mathbf{y}, h_{\mathbf{w}}(X)) + \mathbf{w}^\top \phi_{\boldsymbol{\tau}}(X)\left(h_{\mathbf{w}}(X) - \mathbf{y}\right)$ to the minimization of its upper boundary, which could be obtained by exploring the class label tuple space excluding $\mathbf{y}$, $\mathbf{y}'_l \in \mathcal{Y}/\mathbf{y}$,

$$\begin{aligned} &\Delta(\mathbf{y}, h_{\mathbf{w}}(X)) + \mathbf{w}^\top \phi_{\boldsymbol{\tau}}(X)\left(h_{\mathbf{w}}(X) - \mathbf{y}\right) \\ &\qquad \leq \max_{l:\mathbf{y}'_l \in \mathcal{Y}/\mathbf{y}} \left[\Delta(\mathbf{y}, \mathbf{y}'_l) + \mathbf{w}^\top \phi_{\boldsymbol{\tau}}(X)\left(\mathbf{y}'_l - \mathbf{y}\right)\right] \end{aligned} \quad (8)$$

Thus we can translate the problem in (7) to (9),

$$\min_{\mathbf{w}}\ \left\{\max_{l:\mathbf{y}'_l \in \mathcal{Y}/\mathbf{y}} \left[\Delta(\mathbf{y}, \mathbf{y}'_l) + \mathbf{w}^\top \phi_{\boldsymbol{\tau}}(X)\left(\mathbf{y}'_l - \mathbf{y}\right)\right]\right\}. \quad (9)$$

It could be further relaxed by introducing a nonnegative slack variable ξ to represent the upper boundary, so that the problem could be rewritten as

$$\begin{aligned} &\min_{\mathbf{w},\xi}\ \xi, \\ &s.t.\ \Delta(\mathbf{y}, \mathbf{y}'_l) + \mathbf{w}^\top \phi_{\boldsymbol{\tau}}(X)(\mathbf{y}'_l - \mathbf{y}) \leq \xi, \forall l : \mathbf{y}'_l \in \mathcal{Y}/\mathbf{y}, \\ &\quad \xi \geq 0. \end{aligned} \quad (10)$$

Combining the problems in (2) and (10), and introducing constrains on $\boldsymbol{\tau}$ to prevent negative kernel weights, the following overall optimization problem,

$$\begin{aligned}
&\min_{\mathbf{w},\xi,\boldsymbol{\tau}} \ \frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\xi, \\
&s.t.\ \Delta(\mathbf{y},\mathbf{y}'_l) + \mathbf{w}^\top\phi_{\boldsymbol{\tau}}(X)(\mathbf{y}'_l - \mathbf{y}) \le \xi, l : \mathbf{y}'_l \in \mathcal{Y}/\mathbf{y}, \\
&\quad \xi \ge 0, \ \sum_{m=1}^{M} \tau_m = 1, \tau_m \ge 0, m = 1, \cdots, M.
\end{aligned} \tag{11}$$

where C is a tradeoff parameter.

B. Optimization

To optimize this problem, we give the primal Lagrangian function as follows,

$$\begin{aligned}
&\mathcal{L}(\mathbf{w},\xi,\boldsymbol{\tau},\boldsymbol{\alpha},\beta,\gamma,\boldsymbol{\delta}) = \frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\xi \\
&+ \sum_{l:\mathbf{y}'_l \in \mathcal{Y}/\mathbf{y}} \alpha_l \left(\Delta(\mathbf{y},\mathbf{y}'_l) + \mathbf{w}^\top\phi_{\boldsymbol{\tau}}(X)(\mathbf{y}'_l - \mathbf{y}) - \xi\right) \\
&- \beta\xi - \gamma\left(\sum_{m=1}^{M}\tau_m - 1\right) - \sum_{m=1}^{M}\delta_m\tau_m
\end{aligned} \tag{12}$$

where $\alpha_l \ge 0$, $\beta \ge 0$, $\gamma \ge 0$ and $\delta_m \ge 0$ are the Lagrange multipliers. We argue the following dual optimization problem,

$$\begin{aligned}
&\max_{\boldsymbol{\alpha},\beta,\gamma,\boldsymbol{\delta}} \min_{\mathbf{w},\xi,\boldsymbol{\tau}} \mathcal{L}(\mathbf{w},\xi,\boldsymbol{\tau},\boldsymbol{\alpha},\beta,\gamma,\boldsymbol{\delta}) \\
&\quad s.t.\ \alpha_l \ge 0, l : \mathbf{y}'_l \in \mathcal{Y}/\mathbf{y}, \\
&\qquad \beta \ge 0, \gamma \ge 0, \delta_m \ge 0, m = 1, \cdots, M.
\end{aligned} \tag{13}$$

By setting the digestives of the Lagrange function with regard to $\mathbf{w}$ and ξ to zero, we have

$$\begin{aligned}
&\frac{\partial\mathcal{L}}{\partial\mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{l:\mathbf{y}'_l \in \mathcal{Y}/\mathbf{y}} \alpha_l\phi_{\boldsymbol{\tau}}(X)(\mathbf{y} - \mathbf{y}'_l) \\
&\frac{\partial\mathcal{L}}{\partial\xi} = 0 \Rightarrow C - \sum_{l:\mathbf{y}'_l \in \mathcal{Y}/\mathbf{y}} \alpha_l - \beta = 0 \Rightarrow C \ge \sum_{l:\mathbf{y}'_l \in \mathcal{Y}/\mathbf{y}} \alpha_l.
\end{aligned} \tag{14}$$

By substituting these results and the kernel definition in (1) to (13), we obtain the dual Lagrangian function,

$$\begin{aligned}
&\mathcal{P}(\boldsymbol{\tau},\boldsymbol{\alpha},\gamma,\boldsymbol{\delta}) \\
&= -\frac{1}{2}\sum_{l,k:\mathbf{y}'_l,\mathbf{y}'_k \in \mathcal{Y}/\mathbf{y}} \alpha_l\alpha_k \left((\mathbf{y} - \mathbf{y}'_l)^\top \sum_{m=1}^{M} \tau_m^2 K_m(X,X)(\mathbf{y} - \mathbf{y}'_k)\right) \\
&+ \sum_{l:\mathbf{y}'_l \in \mathcal{Y}/\mathbf{y}} \alpha_l\Delta(\mathbf{y},\mathbf{y}'_l) - \gamma\left(\sum_{m=1}^{M}\tau_m - 1\right) - \sum_{m=1}^{M}\delta_m\tau_m
\end{aligned} \tag{15}$$

This optimization problem is then transformed to

$$\begin{aligned}
&\min_{\boldsymbol{\tau}} \min_{\boldsymbol{\alpha},\gamma,\boldsymbol{\delta}} \mathcal{P}(\boldsymbol{\tau},\boldsymbol{\alpha},\gamma,\boldsymbol{\delta}) \\
&\quad s.t.\ \alpha_l \ge 0, l : \mathbf{y}'_l \in \mathcal{Y}/\mathbf{y},\ C \ge \sum_{l:\mathbf{y}'_l \in \mathcal{Y}/\mathbf{y}} \alpha_l, \\
&\qquad \gamma \ge 0, \delta_m \ge 0, m = 1, \cdots, M.
\end{aligned} \tag{16}$$

To solve this problem, we adopt an alternate optimization strategy. In an iterative algorithm, $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ with its Lagrange multipliers γ and $\boldsymbol{\delta}$ are optimized alternately.

- **Optimizing $\boldsymbol{\alpha}$** By fixing $\boldsymbol{\tau}$ with its Lagrange multipliers γ and $\boldsymbol{\delta}$, and only considering $\boldsymbol{\alpha}$, the optimization problem in (16) is reduced to

$$\begin{aligned}
&\max_{\boldsymbol{\alpha}} \left(-\frac{1}{2}\sum_{l,k:\mathbf{y}'_l,\mathbf{y}'_k \in \mathcal{Y}/\mathbf{y}} \alpha_l\alpha_k \left((\mathbf{y} - \mathbf{y}'_l)^\top K_{\boldsymbol{\tau}}(X,X)(\mathbf{y} - \mathbf{y}'_k)\right)\right. \\
&\qquad \left. + \sum_{l:\mathbf{y}'_l \in \mathcal{Y}/\mathbf{y}} \alpha_l\Delta(\mathbf{y},\mathbf{y}'_l)\right) \\
&s.t. \sum_{l:\mathbf{y}'_l \in \mathcal{Y}/\mathbf{y}} \alpha_l \le C, \alpha_l \ge 0, l : \mathbf{y}'_l \in \mathcal{Y}/\mathbf{y}.
\end{aligned} \tag{17}$$

  This problem can be solved as a quadratic programming problem.
- **Solving $\boldsymbol{\tau}$** By fixing $\boldsymbol{\alpha}$, and only considering $\boldsymbol{\tau}$ and its Lagrange multipliers γ and $\boldsymbol{\delta}$, we have the following problem,

$$\begin{aligned}
&\min_{\boldsymbol{\tau}} \max_{\gamma,\boldsymbol{\delta}} \left\{-\frac{1}{2}\sum_{l,k:\mathbf{y}'_l,\mathbf{y}'_k \in \mathcal{Y}/\mathbf{y}} \left(\alpha_l\alpha_k(\mathbf{y} - \mathbf{y}'_l)^\top\right.\right. \\
&\qquad \left. \times \sum_{m=1}^{M} \tau_m^2 K_m(X,X)(\mathbf{y} - \mathbf{y}'_k)\right) \\
&\qquad \left. -\gamma\left(\sum_{m=1}^{M}\tau_m - 1\right) - \sum_{m=1}^{M}\delta_m\tau_m\right\} \\
&s.t.\ \gamma \ge 0, \delta_m \ge 0, m = 1, \cdots, M.
\end{aligned} \tag{18}$$

  This is the dual form of a constrained quadratic programming problem, and we can solve it as a constrained quadratic programming problem.
- **Updating $\mathcal{Y}/\mathbf{y}$** Moreover, it should be noted that the construction of set $\mathcal{Y}/\mathbf{y}$ is also a problem. To this end, we propose to construct $\mathcal{Y}/\mathbf{y}$ sequentially in the iterative algorithm. We propose to construct $\mathcal{Y}/\mathbf{y}$ by adding one new class label tuple to $\mathcal{Y}/\mathbf{y}$ in each iteration according to updated $\mathbf{w}$ and $\boldsymbol{\tau}$,

$$\begin{aligned}
\mathbf{y}^* = \underset{\mathbf{y}'' \in \{+1,-1\}^n, \mathbf{y}'' \ne \mathbf{y}, \mathbf{y}'' \notin \mathcal{Y}/\mathbf{y}}{\arg\max} &\left\{\Delta(\mathbf{y},\mathbf{y}'') + \right. \\
&\left. \sum_{l:\mathbf{y}'_l \in \mathcal{Y}/\mathbf{y}} \left(\alpha_l(\mathbf{y} - \mathbf{y}'_l)^\top K_{\boldsymbol{\tau}}(X,\mathbf{x}_i)\mathbf{y}''\right)\right\}.
\end{aligned} \tag{19}$$

  where $K_{\boldsymbol{\tau}}(X,\mathbf{x}_i) = [K_{\boldsymbol{\tau}}(\mathbf{x}_1,\mathbf{x}_i), \cdots, K_{\boldsymbol{\tau}}(\mathbf{x}_n,\mathbf{x}_i)]^\top \in \mathbb{R}^{n\times 1}$. Then we can update $\mathcal{Y}/\mathbf{y}$ by adding $\mathbf{y}^*$ to it,

$$\mathcal{Y}/\mathbf{y} \leftarrow \{\mathbf{y}^*\} \cup \mathcal{Y}/\mathbf{y}. \tag{20}$$

C. Algorithm

The iterative multi-kernel learning algorithm to optimize multivariate performance measure is summarized in Algorithm 1.

Algorithm 1 Multi-Kernel Learning algorithm for optimize multivariate Performance measure Optimization (MKLPO).

Input: Training sample feature matrix X, and corresponding class label tuple $\mathbf{y}$;
Initialize $\boldsymbol{\alpha}^0$ and $\boldsymbol{\tau}^0$;
Initialize $\mathcal{Y}/\mathbf{y} = \emptyset$;
for $t = 1, \cdots, T$ **do**
  Obtain a predicted class label tuple $\mathbf{y}^*$ as in (19) by fixing $\boldsymbol{\alpha}^{t-1}$ and $\boldsymbol{\tau}^{t-1}$, and add it to $\mathcal{Y}/\mathbf{y}$ as in (20);
  Update $\boldsymbol{\alpha}^t$ by solving (17) and fixing $\boldsymbol{\tau}^{t-1}$;
  Update $\boldsymbol{\tau}^t$ by solving (18) and fixing $\boldsymbol{\alpha}^t$;
end for
Output: Output the learned $\boldsymbol{\alpha}^T$ and $\boldsymbol{\tau}^T$.

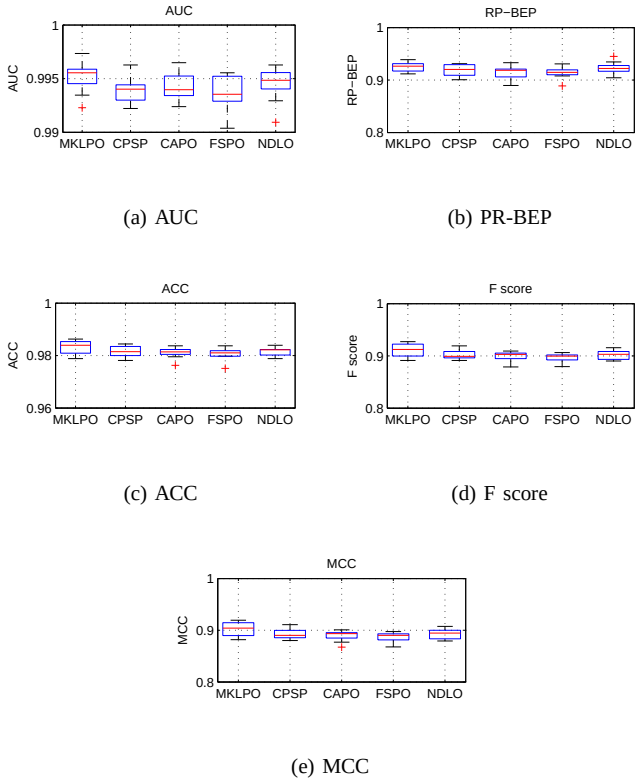
III. Experiments

A. Experiment I: Allergen prediction

In the first experiment, we perform the proposed to the problem of allergen prediction to optimized various prediction performance measures [8].

1) Dataset and protocol: In this experiment, we used a dataset constructed by Dang and Lawrence [8]. This dataset contains 42,977 protein sequences, 3,907 of them are allergens while the remaining 39,070 are non-allergens. To extract feature from each protein sequence, we used the bag-of-words method [9]. Firstly, the amino acid sequence of a protein is broken to some overlapping peptides with a small sliding window, and each peptides is treated as a word. To conduct the experiment, we perform the popular 10-fold cross validation. Various performance measures are considered in this experiment. The multivariate performance measures are optimized on the training set and tested on the test set, including AUC, RP-BEP, ACC, F score and MCC.

2) Results: We compare the proposed multi-kernel learning based multivariate performance measures optimization algorithm agains the original kernel version of SVM^{Perf}, cutting-plane subspace pursuit (CPSP) algorithm [5]. Moreover, three different variations of SVM^{Perf} are also compared as the state-of-the-art multivariate performance measures optimization methods, including the performance measure optimization method by classifier adaptation (CAPO) [10], the feature selection method for multivariate performance measures optimization (FSPO) [11], and the non-decomposable loss functions optimization method (NDLO) [12]. We used these methods to optimize the multivariate performances of AUC of ROC, PR-BEP of recall-precision curve, ACC, F score, and MCC respectively on the training set, and the test them on the test set. The boxplots of the corresponding performance measures of 10-fold cross validations are given in Figure 1. From this figure, we can see clearly that the proposed multi-kernel



(a) AUC (b) PR-BEP

(c) ACC (d) F score

(e) MCC

Fig. 1. Boxplots of optimized multivariate performance measures of 10-fold cross validations of allergen prediction problem.

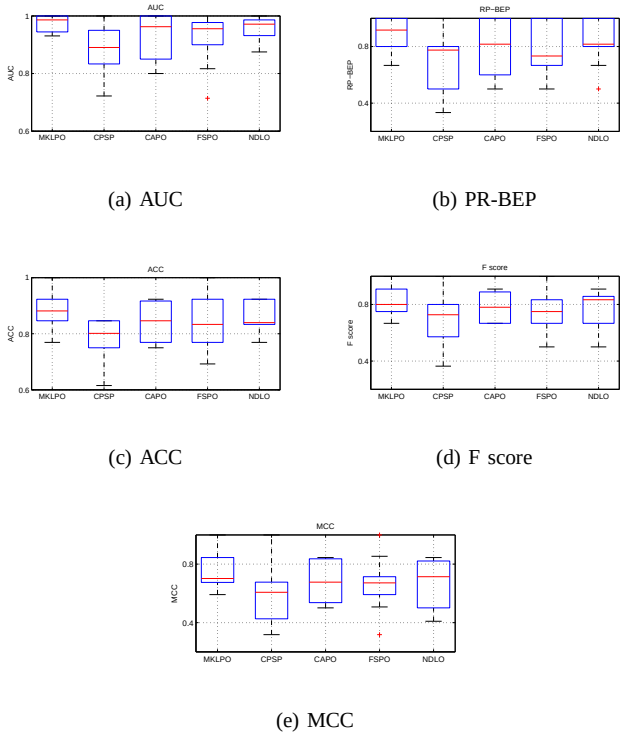
based multivariate performance measure optimization method achieves the best results with regard to different performance measures. Similar phenomenon can be observed in Figure 1(e), and MKLPO is the only algorithm which obtains a higher MCC median value than 0.900. For other performance measures, MKLPO also optimize them to achieve the best performances measures on the test sets. Among the compared algorithms, both CPSP and CAPO are improved by using kernel trickles. However, due to the limitation of single kernel, their performance are not necessarily superior to the linear models, FSPO and NDLO. In most cases, their performances are comparable to each other.

B. Experiment II: Rehabilitative speech treatment assessment

In this experiment, we test the proposed algorithm for the automatic assessment of rehabilitative speech treatment.

1) Dataset and protocol: In this experiment, we use the dataset provided by Tsanas et al. [13]. There are 126 phonations in the data set. A speech expert is employed to assess the phonations, and label them as “acceptable” or “unacceptable”. Among the 126 phonations, 42 is labeled as “acceptable” while the remaining 84 is labeled as “unacceptable”. Each phonation is defined as a data sample in the problem of pattern classification, and ‘acceptable” phonation is defined as positive sample, while “unacceptable” phonation as negative sample. For the purpose of pattern classification, we extract features from each of the phonations. To conduct the experiment,

we also use the 10-fold cross validation. The multivariate performance measures are optimized on the training set and tested on the test set, including AUC, RP-BEP, ACC, F score and MCC.



(a) AUC (b) PR-BEP

(c) ACC (d) F score

(e) MCC

Fig. 2. Boxplots of optimized multivariate performance measures of 10-fold cross validations of rehabilitative speech treatment assessment problem.

2) Results: Fig. 2 shows the boxplots of optimized multivariate performance measures of 10-fold cross validations by using rehabilitative speech treatment assessment data set. As can be seen, our MKLPO algorithm significantly outperforms the other multivariate performance measures optimization algorithms in most cases. The performance difference is larger as the MCC is optimized as the desired multivariate performance measure. The CAPO algorithm outperforms other algorithms in most cases slightly besides the proposed MKLPO algorithm. This result is consistent with the experiment results given in the previous section.

IV. Conclusions and Future Works

Recently a multivariate performance measures optimization method is proposed to estimate a given complex multivariate performance measure as a linear function. This method is based on kernel trick. However, it is difficult to choose a suitable kernel function with its corresponding parameter. To solve this problem, in this paper, we proposed the first multi-kernel learning based algorithm for the problem of optimization of multivariate performance measures. We build a unified objective function for the learning of both multiple kernel weight and classifier parameter for the purpose of multivariate performance measure. An iterative algorithm is developed to optimize the objective function. The experiment results on two different pattern classification problems show that the proposed algorithm outperforms the state-of-the-art multivariate performance measure optimization methods. In the future, we will also explore the potential of using the proposed methods to bioinformatics problems [14], [15], [16], [17], [18], [19], [20], [21], [22], integrated circuit design [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], multiple model big data analysis [33], [34], [35], [36], [37], [38], [39], [40], software and network security [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], and power systems optimization [51], [52]. Moreover, we will also improve the proposed method by regularizing the learning of classifier by graphs [53], [54], [55], [56], [57], [58], [59], [60], [61].

References

[1] T. Joachims, "A support vector method for multivariate performance measures," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 377–384.

[2] K.-R. Mller, S. Mika, G. Rtsch, K. Tsuda, and B. Schlkopf, "An introduction to kernel-based learning algorithms," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, 2001.

[3] Y. Wang, P. Chen, and Y. Jin, "Trajectory planning for an unmanned ground vehicle group using augmented particle swarm optimization in a dynamic environment," in *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*. IEEE, 2009, pp. 4341–4346.

[4] J. J.-Y. Wang, Y. Wang, B.-Y. Jing, and X. Gao, "Regularized maximum correntropy machine," *Neurocomputing*, vol. 160, pp. 85–92, 2015.

[5] T. Joachims and C.-N. Yu, "Sparse kernel svms via cutting-plane training," *Machine Learning*, vol. 76, no. 2-3, pp. 179–193, 2009.

[6] J. J.-Y. Wang, H. Bensmail, and X. Gao, "Feature selection and multi-kernel learning for sparse representation on a manifold," *Neural Networks*, vol. 51, pp. 9–16, 2014.

[7] H. Wang and J. Wang, "An effective image representation method using kernel classification," in *2014 IEEE 26th International Conference on Tools with Artificial Intelligence (ICTAI 2014)*, 2014, pp. 853–858.

[8] H. X. Dang and C. B. Lawrence, "Allerdictor: fast allergen prediction using text classification techniques," *Bioinformatics*, p. btu004, 2014.

[9] J. J.-Y. Wang, H. Bensmail, and X. Gao, "Joint learning and weighting of visual vocabulary for bag-of-feature based tissue classification," *Pattern Recognition*, vol. 46, no. 12, pp. 3249–3255, 2013.

[10] N. Li, I. Tsang, and Z.-H. Zhou, "Efficient optimization of performance measures by classifier adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1370–1382, 2013.

[11] Q. Mao and I.-H. Tsang, "A feature selection method for multivariate performance measures," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 9, pp. 2051–2063, 2013.

[12] M. Ranjbar, T. Lan, Y. Wang, S. N. Robinovitch, Z.-N. Li, and G. Mori, "Optimizing nondecomposable loss functions in structured prediction," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 4, pp. 911–924, 2013.

[13] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, "Objective automatic assessment of rehabilitative speech treatment in parkinson's disease," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 1, pp. 181 – 190, 2014.

[14] H. Chen and E. Ruckenstein, "Formation and degradation of multicomponent multicore micelles: insights from dissipative particle dynamics simulations," *Langmuir*, vol. 29, no. 18, pp. 5428–5434, 2013.

[15] ——, "Nanoparticle aggregation in the presence of a block copolymer," *The Journal of chemical physics*, vol. 131, no. 24, p. 244904, 2009.

[16] J. Wang, Y. Li, Q. Wang, X. You, J. Man, C. Wang, and X. Gao, "Proclusensem: predicting membrane protein types by fusing different modes of pseudo amino acid composition," *Computers in biology and medicine*, vol. 42, no. 5, pp. 564–574, 2012.

[17] F. Yi, I. Moon, and Y. H. Lee, "Three-dimensional counting of morphologically normal human red blood cells via digital holographic microscopy," *Journal of biomedical optics*, vol. 20, no. 1, pp. 016 005–016 005, 2015.

[18] Y. Wang, H.-C. Han, J. Y. Yang, M. L. Lindsey, and Y. Jin, "A conceptual cellular interaction model of left ventricular remodelling post-mi: dynamic network with exit-entry competition strategy," *BMC systems biology*, vol. 4, no. Suppl 1, p. S5, 2010.

[19] B. Peng, Y. Liu, Y. Zhou, L. Yang, G. Zhang, and Y. Liu, "Modeling nanoparticle targeting to a vascular surface in shear flow through diffusive particle dynamics," *Nanoscale Research Letters*, vol. 10, no. 1, p. 235, 2015.

[20] Y. Liu, J. Yang, Y. Zhou, and J. Hu, "Structure design of vascular stents," *Multiscale simulations and mechanics of biological materials*, pp. 301–317, 2013.

[21] Y. Zhou, W. Hu, B. Peng, and Y. Liu, "Biomarker binding on an antibody-functionalized biosensor surface: the influence of surface properties, electric field, and coating density," *The Journal of Physical Chemistry C*, vol. 118, no. 26, pp. 14 586–14 594, 2014.

[22] J. Luo and A. Brodsky, "Piecewise surface regression modeling in intelligent decision guidance system," in *Intelligent Decision Technologies*, 2011, pp. 223–235.

[23] L. Zhang, M. Gunji, S. Thombare, and P. C. McIntyre, "Eot scaling of on germanium pmosfets and impact of gate metal selection," *Electron Device Letters, IEEE*, vol. 34, no. 6, pp. 732–734, 2013.

[24] L. Zhang, J. Zhuge, Y. Wang, R. Huang, C. Liu, D. Wu, Z. Kang, D.-W. Kim, and D. Park, "New insights into oxide traps characterization in gate-all-around nanowire transistors with tin metal gates based on combined i g-i d rts technique," in *VLSI Technology, 2009 Symposium on*. IEEE, 2009, pp. 46–47.

[25] L. Zhang, R. Wang, J. Zhuge, R. Huang, D.-W. Kim, D. Park, and Y. Wang, "Impacts of non-negligible electron trapping/detrapping on the nbti characteristics in silicon nanowire transistors with tin metal gates," in *2008 IEEE International Electron Devices Meeting*, 2008, pp. 1–4.

[26] L. Zhang, M. Gungi, and P. C. McIntyre, "Germanium channel p-mosfet with tio2/al2o3 bilayer high-k gate stacks and solutions for metal/tio2 interface stability," in *Silicon-Germanium Technology and Device Meeting (ISTDM), 2012 International*. IEEE, 2012, pp. 1–2.

[27] L. Zhang, Z. Kang, R. Wang, and R. Huang, "A comprehensive study on schottky barrier nanowire transistors (sb-nwts): Principle, physical limits and parameter fluctuations," in *Solid-State and Integrated-Circuit Technology, 2008. ICSICT 2008. 9th International Conference on*. IEEE, 2008, pp. 157–160.

[28] J. Zhuge, L. Zhang, R. Wang, R. Huang, D.-W. Kim, D. Park, and Y. Wang, "Random telegraph signal noise in gate-all-around silicon nanowire transistors featuring coulomb-blockade characteristics," *Applied Physics Letters*, vol. 94, no. 8, p. 083503, 2009.

[29] Z. Kang, L. Zhang, R. Wang, and R. Huang, "Investigations on the physical limitation and electrostatic improvement of a gate-all-around silicon nanowire transistor with schottky barrier source/drain," *Semiconductor Science and Technology*, vol. 24, no. 10, p. 105001, 2009.

[30] R. Wang, R. Huang, L. Zhang, H. Liu, D.-W. Kim, D. Park, and Y. Wang, "Experimental investigations on channel backscattering characteristics of gate-all-around silicon nanowire transistors from top-down approach," *Applied Physics Letters*, vol. 93, no. 8, p. 083513, 2008.

[31] F. Zhang, Y. Zhang, and J. D. Bakos, "Accelerating frequent itemset mining on graphics processing units," *The Journal of Supercomputing*, vol. 66, no. 1, pp. 94–117, 2013.

[32] Y. Gao, F. Zhang, and J. D. Bakos, "Sparse matrix-vector multiply on the keystone ii digital signal processor," in *High Performance Extreme Computing Conference (HPEC), 2014 IEEE*, 2014, pp. 1–6.

[33] J. Wang, Y. Zhou, K. Duan, J. J.-Y. Wang, and H. Bensmail, "Supervised cross-modal factor analysis for multiple modal data classification," in *Systems, Man and Cybernetics (SMC), 2015 IEEE International Conference on*. IEEE, 2015.

[34] T. Li, X. Zhou, K. Brandstatter, D. Zhao, K. Wang, A. Rajendran, Z. Zhang, and I. Raicu, "Zht: A light-weight reliable persistent dynamic scalable zero-hop distributed hash table," in *Parallel & Distributed Processing (IPDPS), 2013 IEEE 27th International Symposium on*. IEEE, 2013, pp. 775–787.

[35] T. Li, X. Zhou, K. Brandstatter, and I. Raicu, "Distributed key-value store on hpc and cloud systems," in *2nd Greater Chicago Area System Research Workshop (GCASR)*. Citeseer, 2013.

[36] K. Wang, A. Kulkarni, X. Zhou, M. Lang, and I. Raicu, "Using simulation to explore distributed key-value stores for exascale system services," in *2nd Greater Chicago Area System Research Workshop (GCASR)*, 2013.

[37] K. Wang, M. Lang, X. Zhou, B. McClelland, K. Qiao, and I. Raicu, "Towards scalable distributed workload manager with monitoring-based weakly consistent resource stealing," in *HPDC '15 Proceedings of the 24th International Symposium on High-Performance Parallel and Distributed Computing*, 2015, pp. 219–222.

[38] K. Wang, X. Zhou, H. Chen, M. Lang, and I. Raicu, "Next generation job management systems for extreme-scale ensemble computing," in *Proceedings of the 23rd international symposium on High-performance parallel and distributed computing*, 2014, pp. 111–114.

[39] K. Wang, X. Zhou, T. Li, D. Zhao, M. Lang, and I. Raicu, "Optimizing load balancing and data-locality with data-aware scheduling," in *Big Data (Big Data), 2014 IEEE International Conference on*, 2014, pp. 119–128.

[40] D. Zhao, Z. Zhang, X. Zhou, T. Li, K. Wang, D. Kimpe, P. Carns, R. Ross, and I. Raicu, "Fusionfs: Toward supporting data-intensive scientific applications on extreme-scale high-performance computing systems," in *Big Data (Big Data), 2014 IEEE International Conference on*, 2014, pp. 61–70.

[41] Q. Sun, P. Wu, Y. Wu, M. Guo, and J. Lu, "Unsupervised multi-level non-negative matrix factorization model: Binary data case," *Journal of Information Security*, vol. 3, no. 04, p. 245, 2012.

[42] Q. Sun, W. Yu, N. Kochurov, Q. Hao, and F. Hu, "A multi-agent-based intelligent sensor and actuator network design for smart house and home automation," *Journal of Sensor and Actuator Networks*, vol. 2, no. 3, pp. 557–588, 2013.

[43] Q. Sun, F. Hu, and Q. Hao, "Human movement modeling and activity perception based on fiber-optic sensing system," *Human-Machine Systems, IEEE Transactions on*, vol. 44, no. 6, pp. 743–754, 2014.

[44] ——, "Mobile target scenario recognition via low-cost pyroelectric sensing system: Toward a context-enhanced accurate identification," *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, vol. 44, no. 3, pp. 375–384, 2014.

[45] F. Yi, I. Moon, and Y. H. Lee, "A multispectral photon-counting double random phase encoding scheme for image authentication," *Sensors*, vol. 14, no. 5, pp. 8877–8894, 2014.

[46] S. Zhang, D. Caragea, and X. Ou, "An empirical study on using the national vulnerability database to predict software vulnerabilities," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6860 LNCS, no. PART 1, pp. 217–231, 2011.

[47] S. Zhang, X. Zhang, and X. Ou, "After we knew it: empirical study and modeling of cost-effectiveness of exploiting prevalent known vulnerabilities across iaas cloud," in *Proceedings of the 9th ACM symposium on Information, computer and communications security*, 2014, pp. 317–328.

[48] S. Zhang, X. Ou, and J. Homer, "Effective network vulnerability assessment through model abstraction," in *Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 2011, pp. 17–34.

[49] H. Huang, S. Zhang, X. Ou, A. Prakash, and K. Sakallah, "Distilling critical attack graph surface iteratively through minimum-cost sat solving," in *Proceedings of the 27th Annual Computer Security Applications Conference*. ACM, 2011, pp. 31–40.

[50] R. Zhuang, S. Zhang, A. Bardas, S. A. DeLoach, X. Ou, and A. Singhal, "Investigating the application of moving target defenses to network security," in *Resilient Control Systems (ISRCS), 2013 6th International Symposium on*, 2013, pp. 162–169.

[51] L. Che and M. Shahidehpour, "Dc microgrids: Economic operation and enhancement of resilience by hierarchical control," *IEEE Transactions Smart Grid*, vol. 5, no. 5, pp. 2517–2526, 2014.

[52] L. Che, M. Khodayar, and M. Shahidehpour, "Only connect: Microgrids for distribution system restoration," *IEEE Power and Energy Magazine*, vol. 12, no. 1, pp. 70–81, 2014.

[53] J.-Y. Wang, I. Almasri, and X. Gao, "Adaptive graph regularized nonnegative matrix factorization via feature selection," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, 2012, pp. 963–966.

[54] J. J.-Y. Wang, H. Bensmail, and X. Gao, "Multiple graph regularized protein domain ranking," *BMC bioinformatics*, vol. 13, no. 1, p. 307, 2012.

[55] W. Shen, J. Wang, and S. Ma, "Doubly regularized portfolio with risk minimization," in *Proceedings of the National Conference on Artificial Intelligence*, vol. 2, 2014, pp. 1286–1292.

[56] W. Shen and J. Wang, "Transaction costs-aware portfolio optimization via fast lowner-john ellipsoid approximation," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 1854 – 1860.

[57] L. Wang and J. Isberg, "Nonlinear passive control of a wave energy converter subject to constraints in irregular waves," *Energies*, vol. 8, no. 7, pp. 6528–6542, 2015.

[58] L. Wang, J. Engström, M. Göteman, and J. Isberg, "Constrained optimal control of a point absorber wave energy converter with linear generator," *Journal of Renewable and Sustainable Energy*, vol. 7, no. 4, p. 043127, 2015.

[59] X. Liu, J. Wang, M. Yin, B. Edwards, and P. Xu, "Supervised learning of sparse context reconstruction coefficients for data representation and classification," *Neural Computing and Applications*, 2015.

[60] J. Wang, Y. Zhou, M. Yin, S. Chen, and B. Edwards, "Representing data by sparse combination of contextual data points for classification," in *Advances in Neural Networks–ISNN 2015*. Springer, 2015.

[61] J. Wang, Y. Zhou, H. Wang, X. Yang, F. Yang, and A. Peterson, "Image tag completion by local learning," in *Advances in Neural Networks–ISNN 2015*. Springer, 2015.